PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The Structure of Some Proteins as Revealed by an X-Ray Scattering Method

U. W. Arndt and D. P. Riley

Phil. Trans. R. Soc. Lond. A 1955 **247**, 409-439
doi: 10.1098/rsta.1955.0002

THE STRUCTURE OF SOME PROTEINS AS REVEALED BY AN X-RAY SCATTERING METHOD

By U. W. ARNDT and D. P. RILEY*

Royal Institution, Albemarle Street, London, W. 1

(*Communicated by L. Pauling, For.Mem.R.S.—Received* 28 *July* 1954)

CONTENTS

The intensity of scattering of monochromatic X-rays by isotropic specimens of thirty proteins (or protein fractions) and two synthetic polypeptides has been measured by means of a proportional-counter diffractometer. The curves fall into only three essentially different types, called α, β and γ. The α-proteins are the most numerous and some twenty have been examined; all the intensity curves are very similar and suggest that a single polypeptide-chain configuration is common to all as the predominant constituent. This predominant configuration is considered to be best represented by

* Now at Queen Mary College, London, E. 1.

one of the variants of the α-helix of Pauling & Corey (1951 *a*) and Pauling, Corey & Branson (1951) to which the term α_1-helix is given. The method of interpretation used was to calculate the intensity of scattering that would be produced by independent polypeptide chains in random orientation, various regular configurations being assigned to them in accordance with existing hypotheses. In this way, the scattering characteristics of the following helical chain configurations were derived: α, γ, π, 4_{13}, 3_{10}, 3_8 and 2_7; in the first five mentioned, the two alternative positions of the β-carbon atom were considered. Comparison with the observed curves for α-proteins showed that the curve for the α_1-helix was in best agreement with experiment, and with the synthetic polypeptides the correlation was particularly convincing. It was further shown that the several helices, or parts of helices, that constitute a whole protein molecule are usually packed together compactly in a nearly parallel fashion; in certain cases, a more pronounced skewness is indicated.

1. Introduction

The first X-ray diffraction patterns obtained from protein crystals were of the powder type, but this method was almost immediately abandoned in favour of the single-crystal method with its greater wealth of data. This movement was made possible by the demonstration by Bernal & Crowfoot (1934) that if a protein crystal (in this case, pepsin) were kept immersed in its mother liquor of crystallization, the diffraction pattern obtained was very rich and corresponded to a high degree of perfection in the crystal structure. It was therefore tempting to apply the usual methods of crystallographic structure analysis, or to devise new ones, with the ultimate aim of elucidating the detailed internal structure of a protein molecule. Since that time, a great deal of work along these lines has been done by some of our most skilful analysts and many and ingenious have been the procedures employed. In spite of this effort, little certain knowledge about the actual structural chemistry of proteins has so far been gathered by this method, although a considerable amount of crystallographic information of great potential usefulness has been accumulated.

An alternative approach to the problem was adopted from 1930 onwards by Astbury who, in the main, confined his attention to the fibrous proteins (see, for example, Astbury 1931, 1940, 1953). He and his co-workers examined a large number of these and were able to establish the existence of only a limited number of structural types. Whereas Meyer & Mark (1928) had assigned a fully extended polypeptide chain structure to silk, Astbury's extensive surveys showed that, in other proteins, the polypeptide chains must be coiled or folded and that there must be at least two different ways in which this can be done, corresponding to keratin and collagen respectively.

Although there are only a limited number of diffraction data on an X-ray fibre photograph, it is relatively easy to relate them to certain features in the molecule itself and, by comparing a number of photographs from different specimens, to draw conclusions from them regarding protein fibres in general. It is difficult from a single photograph, without making any chemical assumptions, to derive any useful knowledge at all. This applies *a fortiori* to powder diagrams as these do not even have the benefit of preferred molecular orientation.

At first sight, then, it would appear that single-crystal analysis would be the most powerful diffraction procedure to use and the powder method the least, and that this would be even more so if the powders consisted of amorphous or imperfectly crystalline material. This conclusion is generally true, particularly if the strictly deductive procedure is adopted in which the fewest assumptions (ideally, none) are made concerning the

structure of the molecule. It is doubtful whether such a pure method can ever completely succeed in as complex a structure as a protein. Inevitably, certain hypotheses will be advanced and will be compared with the experimental data; the crystallographic facts will not be considered in strict isolation. This, after all, is the basis of the well-established 'trial and error' approach to crystal structure analysis which has only quite recently been largely supplanted by the intensive use of Fourier series in 'direct' methods. In short, it is not until the single-crystal Fourier method has produced electron-density maps in which the separate atoms in the structure are clearly recognizable that it will have succeeded as a direct method; until this degree of resolution is reached there will still remain the problem of interpreting the maps constituting a coarsely resolved image of the structure.

The single-crystal method is not the most economical way of arriving at an approximate solution of the molecular structure, although it may be the only one which has the potentiality of achieving a complete and exact determination. We were therefore led to seek an easier and quicker method of testing the various hypotheses of protein structure that have been advanced in recent years and to re-examine the powder, or amorphous scattering, method as a mean to this end. A complete structural analysis has not been aimed at; this would be an impossibility with the methods used. It should, however, be feasible to determine the nature of the predominant configuration of the polypeptide chain in a protein molecule and this we have attempted to do. The chain configuration represents the master-pattern or essential core of the molecule, of which the whole structure is an intricate elaboration. Its determination has long been considered the central problem of protein structural chemistry. Knowledge of the configuration would greatly assist a detailed understanding of the chemistry and biological function of the molecule and provide a solid base for reasoning from the results of amino-acid sequence analyses. While good evidence has been adduced in favour of one definite configurational hypothesis from fibre photographs of synthetic polypeptides and of certain fibrous proteins, the position as regards the large body of corpuscular proteins has remained obscure. The present work has had as its principal object the elucidation of the nature of polypeptide chain configuration in corpuscular proteins.

2. General description of the method

The ideal diffraction experiment would record the spatial distribution of intensity scattered by a single molecule as a function of its orientation relative to the primary beam, and would completely exclude confusing intermolecular interference effects. While it is experimentally very difficult to derive the molecular intensity function with this degree of explicitness, it is simple to measure its average value over all orientations. This is the basis of those methods, using the diffraction of either X-rays or electrons, which determine the structure of simple molecules in the gas or vapour state. In a gas, each molecule is effectively an independent scatterer, and the observed intensity is the sum of the intensities given by many molecules in random orientation. Debye (1915, 1930) was the pioneer of the method and it has been extensively applied by Pauling & Brockway (1934) and others, in its electron-diffraction aspect, to the study of simple molecules. In particular, Warren and his collaborators (e.g. Warren & Gingrich 1934) have investigated the structure of some relatively simple amorphous solids by similar X-ray means.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

With a protein, a dilute solution free from aggregation effects would correspond to the gas state. The effects of intermolecular interferences are, however, relatively unimportant with large molecules and are localized in the low-angle region of scattering. It is therefore permissible to examine proteins in the form of isotropic solid specimens and to consider that the major part of the intensity curve, which is due to intramolecular interferences, is not appreciably influenced by the state of aggregation. A powdered sample tightly compressed into a block constitutes an excellent specimen; it is immaterial whether the powder is truly amorphous or crystalline. The essential is that the molecules contained in the irradiated volume possess random orientation.

The intensity of monochromatic X-rays scattered by such a specimen is measured over a wide angular range. Two procedures are then possible. One is to compare this intensity curve with equivalent curves calculated, using well-known relations, from a number of postulated molecular models. The other is to derive, by Fourier transformation of the experimental data, a radial distribution curve characteristic of the average atomic environment in the molecule and to use this as the basis of comparison with theory. In a previous note (Riley & Arndt 1952), we used the latter procedure, which has also been applied by Bjørnhaug, Ellefsen & Tønnesen (1954) to the study of high polymers. In the light of experience acquired during the present work, we have been led to prefer the alternative method of intensity comparison as being both more delicate and more reliable. The relative merits of the two procedures are further discussed in §4 (*d*) (see also Riley & Arndt 1954).

No novelty is claimed for the method as such. Its application to such complex structures as proteins, however, raises a number of new problems both on the experimental and theoretical sides. In, at least partly, solving these, the work reported here will, we hope, serve to establish the amorphous scattering method as one meriting serious consideration in the high-polymer field generally. The method is not capable of leading directly, as can certain single-crystal Fourier techniques, to a detailed picture of the molecule. It is essentially a simple and rapid means of testing predicted structures, which can be highly complicated, and is adequately sensitive for its purpose (see §4 (*d*)).

In order to compare theory with experiment absolutely, it would be necessary that any structural hypothesis be stated in explicit form giving the co-ordinates of every atom in the molecule. This has not yet been done for the whole of a protein molecule but, if it were, there would be no formal difficulty in making the comparison. Hypotheses are available based on the well-established chemical fact that a protein is constituted of a small number of main polypeptide chains to which are attached at regular intervals a large number of relatively small atomic groupings, the side-chains or *R*-groups. Various schemes for folding or coiling polypeptide chains have been suggested and a number of them give with some exactness the co-ordinates of the atoms C_α, C_β, C′, O and N in a succession of amino-acid residues (see figure 1). It is a simple, albeit laborious, matter to calculate the intensity that would be given by these folded polypeptide chains acting as independent scatterers and to compare it with experiment. The calculations necessarily neglect the contributions of the *R*′-atoms, or non-C_β-atoms in the *R*-groups. This omission does not invalidate the comparisons (see §§4 (*e*) and 8 (*d*)), but precludes any possibility of an exact correspondence between calculation and experiment. Its effect is, however, readily apparent and recogniz-

able. The result of chain-to-chain packing interferences is easily distinguishable in the experimental intensity curves and has little influence on the part due to intra-chain structure (see §§ 4 (f) and 9).

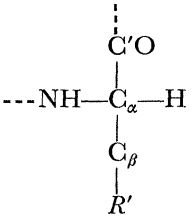

FIGURE 1. An amino-acid residue in the L-configuration found in natural proteins, drawn according to the usual convention. C_α is in the plane of the paper; the bonds to C′O and C_β are below, and those to NH and H above, this plane.

3. FUNDAMENTAL RELATIONS

(a) *Scattering by independent molecular units*

The basic relation on which all calculations are based is due to Debye (1930), and expresses the average intensity per molecule scattered coherently by an assemblage of independent molecules in random orientation:

$$I(s) = \sum_i \sum_j f_i(s)\, f_j(s) \frac{\sin sl_{ij}}{sl_{ij}}. \tag{1}$$

Throughout this paper, the angle variable is defined as

$$s = 4\pi \frac{\sin \theta}{\lambda}, \tag{2}$$

where 2θ is the angle of scattering and λ the wave-length of the X-radiation employed. The double summation embraces all pairs of atoms in the molecule, each pair being counted twice except for the self-identity cases where $i = j$. The symbol l_{ij} refers to the distance between the two atoms, i and j, constituting a pair, while $f_i(s)\, f_j(s)$ represents the product of their scattering, or form, factors.

Debye's expression can be applied to any rigid assembly of atoms acting as an independent scattering unit. In the present investigation, the independent molecular unit is considered to be the regular core of the polypeptide chain as described in the previous section.

It is desirable to separate out the portion of the scattered intensity that is due to inter-atomic interferences, as this is the only part that is sensitive to the nature of the chain configuration. The total intensity $I(s)$ can be split into two parts:

$$I(s) = A(s) + B(s). \tag{3}$$

$A(s)$ corresponds to the cases where $i = j$ in (1) and represents the intensity due to the atoms present scattering independently of each other. Its value is given by

$$A(s) = \sum_i f_i^2(s), \tag{4}$$

the sum of the squares of the atomic scattering factors of all the atoms in the structure. The interference scattering, where $i \neq j$, is represented by

$$B(s) = \sum_i^{i \neq j} \sum_j f_i(s)\, f_j(s) \frac{\sin sl_{ij}}{sl_{ij}}. \tag{5}$$

For organic substances it is possible to effect a considerable simplification by making the assumption that the atomic scattering factors for carbon, nitrogen and oxygen all have the same dependence on the angle variable s and differ only in magnitude, as defined by the atomic number, Z, thus:

$$f_z(s) = Zf(s), \tag{6}$$

where $f(s)$, the normalized factor, is identical for each atom. It is now possible to bring $f^2(s)$ outside the summations in (5) and to write

$$\begin{aligned} i(s) &= B(s)/A(s) \\ &= \nu^{-1} \sum_{i}^{i \neq j} \sum_{j} W_{ij} \frac{\sin sl_{ij}}{sl_{ij}}, \end{aligned} \tag{7}$$

where $\nu = \Sigma Z^2$, the sum of the squares of the atomic numbers of all the atoms concerned, and $W_{ij} = Z_i Z_j$, the product of atomic numbers in the pair of atoms i, j. We shall term $i(s)$ the *interference intensity function*. As it is the ratio of two intensities, it has the dimensions of a pure number and is independent of the nature of $f(s)$, which does not therefore require definition. The interference intensity function provides a convenient means of comparing theory with experiment, as an equivalent function is derivable from the experimental data, as discussed in §§ 6 (*b*) and 7.

(*b*) *Computing procedures*

Equation (7) is not in a suitable form for routine calculation owing to the very large number of different values of $(\sin x)/x$ involved. While with small molecules the use of $(\sin x)/x$ tables is obligatory, with a high polymer advantage can be taken of its size to rewrite (7) as the equivalent Fourier integral

$$\nu s i'(s) = \int_0^\infty \frac{\phi(r)}{r} \sin sr \, \mathrm{d}r, \tag{8}$$

in which the smoothed continuous radial distribution function $\phi(r)$ replaces the discontinuous point function $W_{ij}(l_{ij})$ in (7). The latest Beevers–Lipson strips (Beevers 1952) provide a convenient means of evaluating such integrals.

The radial distribution function is defined as

$$\phi(r) = \int_0^\infty H(l) \, \sigma(r-l) \, \mathrm{d}l, \tag{9}$$

where $\sigma(x)$ is a smoothing function folded into the discontinuous histogram $H(l)$ representative of the distribution of W_{ij} in terms of l_{ij},

$$H(l) = \sum_{l}^{l+\Delta l} W_{ij}(l_{ij}). \tag{10}$$

The intensity function in (8) has been marked with a prime because it includes an effect due to the smoothing function. This can be eliminated by dividing $i'(s)$ by the Fourier transform of $\sigma(x)$, whose exact form is therefore merely a matter of convenience. The justification for this procedure lies in the theorem which states that the Fourier transform of the convolution (or fold) of two functions is equal to the product of the Fourier transforms of the separate functions (see, for instance, Titchmarsh (1937)). $i(s)$, as

finally derived via (8) in this way, differs from the function defined by (7) only by as much as the histogram (10) is an inexact representation of the molecule. By choosing Δl to be 0·1Å, the deviation can be made negligible.

4. Scattering characteristics of some polypeptide chain configurations

(a) *Configurations examined*

Calculations have been carried out for the following helical models of a folded polypeptide chain:

$$\alpha,\ \gamma,\ \pi,\ 4_{13},\ 3_{10},\ 3_8,\ 2_7\,b.$$

For each of the first five helices listed, the two alternative positions of the C_β atoms have been considered, corresponding to a change of hand of the helix if the L-C_α configuration is preserved. As a record of the exact atomic co-ordinates involved is desirable, references to the sources used are given in table 1. Each helix is described according to the nomenclature of Bragg, Kendrew & Perutz (1950), in which S_R represents an S-fold helix containing R atoms in the hydrogen-bonded ring.

Table 1. Helical chain configurations examined

S_R	C_β positions	source of co-ordinates	remarks
$3{\cdot}6_{13}$ (α)	1 (α_2) 2 (α_1)	Pauling & Corey (1951*a*, p. 240)	—
$5{\cdot}1_{17}$ (γ)	1 (γ_2) 2 (γ_1)	Pauling & Corey (1951*a*, p. 238)	per-residue repeat distance adjusted to 0·96 Å
$4{\cdot}3_{16}$ (π)	1 (π_2) 2 (π_1)	Low & Grenville-Wells (1953, p. 799)	C_β co-ordinates sent privately
$4{\cdot}0_{13}$	1 2	Bragg *et al.* (1950, p. 339)	—
$3{\cdot}0_{10}$	1 2	Donohue (1953, p. 472)	—
$3{\cdot}0_8$	—	Bragg *et al.* (1950, p. 338)	—
$2{\cdot}0_7$	*b*	Bragg *et al.* (1950, p. 333)	—

(b) *Radial distribution curves*

Figure 2 assembles the radial distribution curves $\phi(r)$ computed for the model structures mentioned above. In all cases the smoothing function $\sigma(x)$ employed is depicted in figure 3 and corresponds roughly to the electronic distribution around a typical atom. As can be seen, this function approximates to the Gaussian $e^{-9{\cdot}5x^2}$ which has the same half-peak width.

It is clear that the general trend of the curves will depend on the length of the helix, i.e. the number of amino-acid residues introduced into the calculation. The effect is exemplified in figure 2, the number of residues involved being marked for each curve. An infinitely long chain gives rise to a radial distribution function which, at high values of r, oscillates asymptotically about a constant finite value; a chain of finite length gives, after the initial region, a curve that descends steadily until it reaches zero at a value of r approximately equal to its combined length and diameter.

As would be expected, the initial part of all the curves is virtually the same, corresponding to nearest and next-nearest atomic neighbours (peaks at 1·5 and 2·6 Å). The region that is most sensitive to chain configuration lies between about $r = 4$ Å and $r = 10$ Å. The prominent peak, between 4 and 5 Å, that occurs in most cases, is related to the average diameter of the helix.

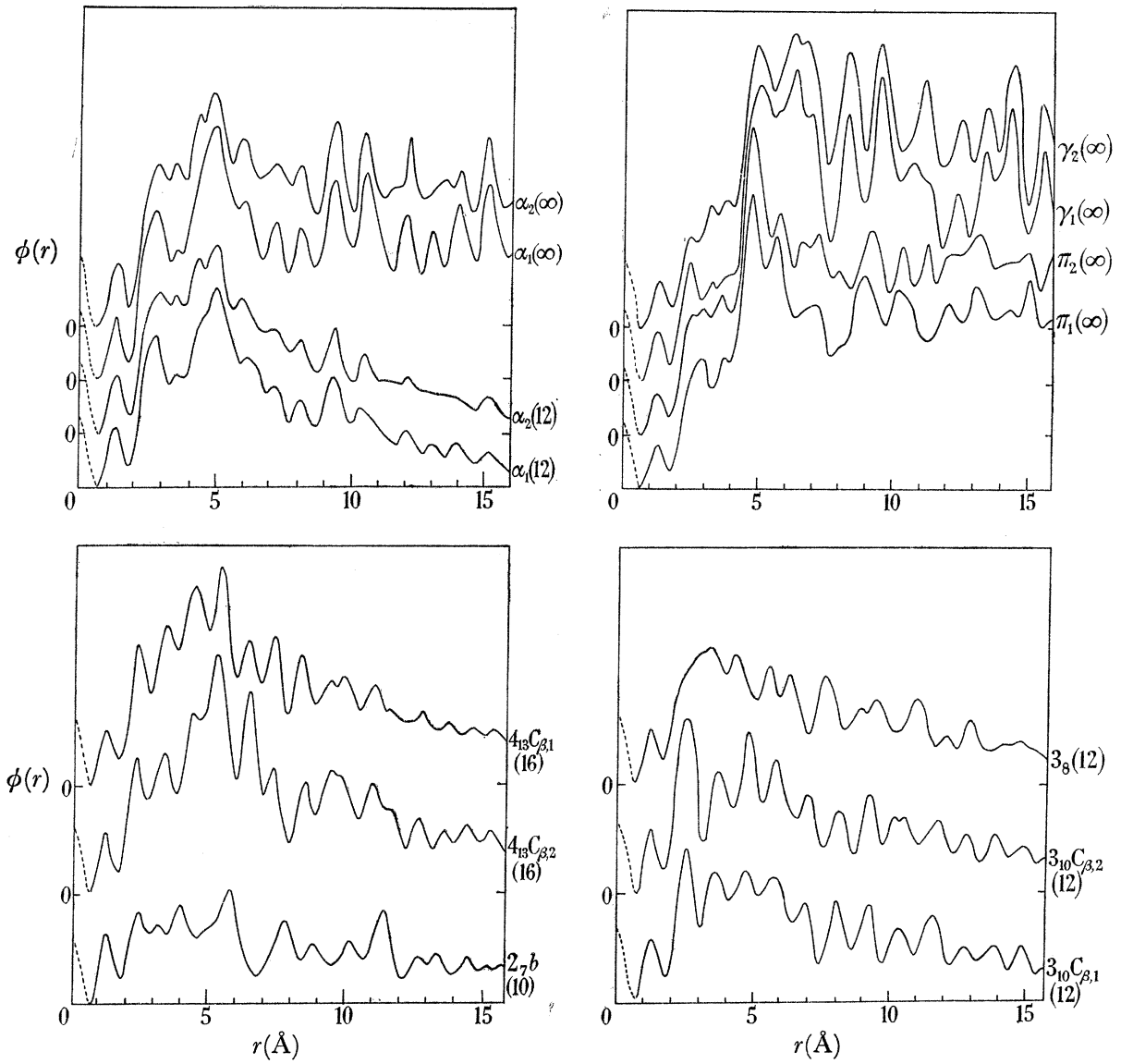

FIGURE 2. Radial distribution curves, $\phi(r)$, derived for various helical configurations of a polypeptide chain. The number of amino-acid residues included in the calculations is shown in brackets.

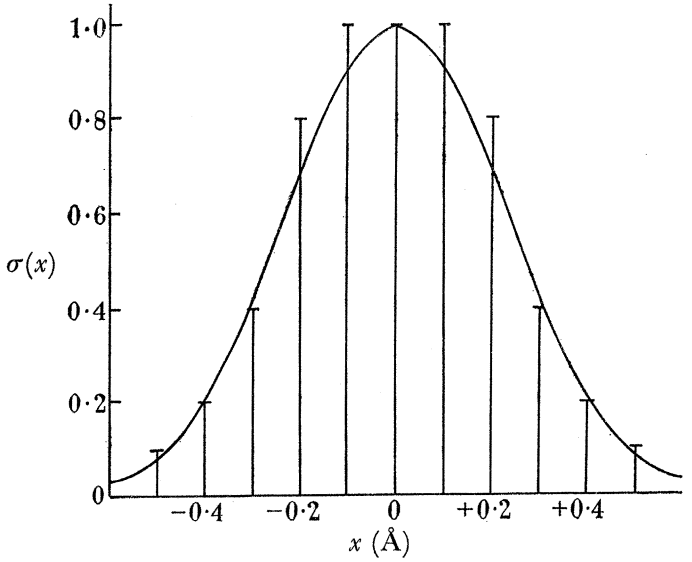

FIGURE 3. The smoothing function, shown by vertical lines, employed in the derivation of radial distribution curves from the atomic co-ordinates of a model. The smooth curve represents the Gaussian $e^{-9\cdot5x^2}$.

(c) *Interference intensity functions*

The interference intensity functions $i(s)$ for the infinitely long α, γ, π and 4_{13} helices are presented in figure 4, the two sets of curves referring to the two alternative positions of the C_β-atoms. The equivalent results for the 3_{10}, 3_8 and 2_7 helices are given in figure 5.

The function $i(s)$ was computed according to the description given in §3 but was modified in one small particular in order to make it more realistic. The modification takes into account the effect of thermal agitation. While James (1932) has shown that thermal movements of covalently bonded atoms make only a negligible impression on the X-ray scattering curves from free molecules, this cannot be assumed to be so for molecules in which the much weaker hydrogen-bonding plays an essential role. The polypeptide chain is held in its helical configuration by hydrogen bonds, and the principal effect of thermal agitation will therefore be a sort of concertina motion along the length of the helix. We have estimated, from the bond energies and atomic masses involved, that this concertina motion will lead to a temperature factor of approximately $e^{-0{\cdot}013s^2}$ by which $i(s)$, calculated from a rigid model, should be multiplied. All the theoretical $i(s)$ curves are on an absolute scale and are strictly comparable.

(d) *Sensitivity and relative merits of the two procedures*

Either radial distribution or $i(s)$ curves can be used as the basis of comparison of theory with experiment. Quite apart from the experimental side, it is clear from an examination of the two types for a number of molecular models that the intensity function is to be preferred. The general similarity between the radial distribution curves for the two variants of the α, π and 4_{13} helices, for instance, would make distinction between them an uncertain matter. Nevertheless, the radial distribution curve method is well suited to rapid diagnosis and is a convenient means of narrowing down the field of choice to reasonable dimensions for subsequent detailed analysis of intensity curves.

In a formal sense, the sensitivity of the two methods is identical since they are only alternative manifestations of the same reality. In practice, the $i(s)$ method is the more sensitive because differences due to configurational changes tend to be localized in one part of the curve and are, thereby, more readily apparent. This is clearly demonstrated by reference to figure 4. The portions of the curves from $s \simeq 3{\cdot}5$ outwards hardly alter when the position of the C_β atom is changed in a given helix; the initial parts, on the other hand, are substantially modified. By close attention to the region between $s \simeq 1$ and $s \simeq 3{\cdot}5$, it is feasible, in principle, to distinguish even between two closely similar variants of the same helix, differing only in the angle of attachment of the side-chains and possessing identical cores of C_α, C′, N and O atoms. The equivalent differences in the case of the corresponding radial distribution curves (figure 2) are more evenly spread out and are consequently more difficult to perceive.

It is evident that homometric structures, defined (Patterson 1944) as possessing the same interatomic distances terminated by similar pairs of like or different atoms, will give rise to identical scattering and radial distribution curves. In a completely general sense, the potentiality of occurrence of nearly homometric structures is considerable when large numbers of similar atoms are involved.

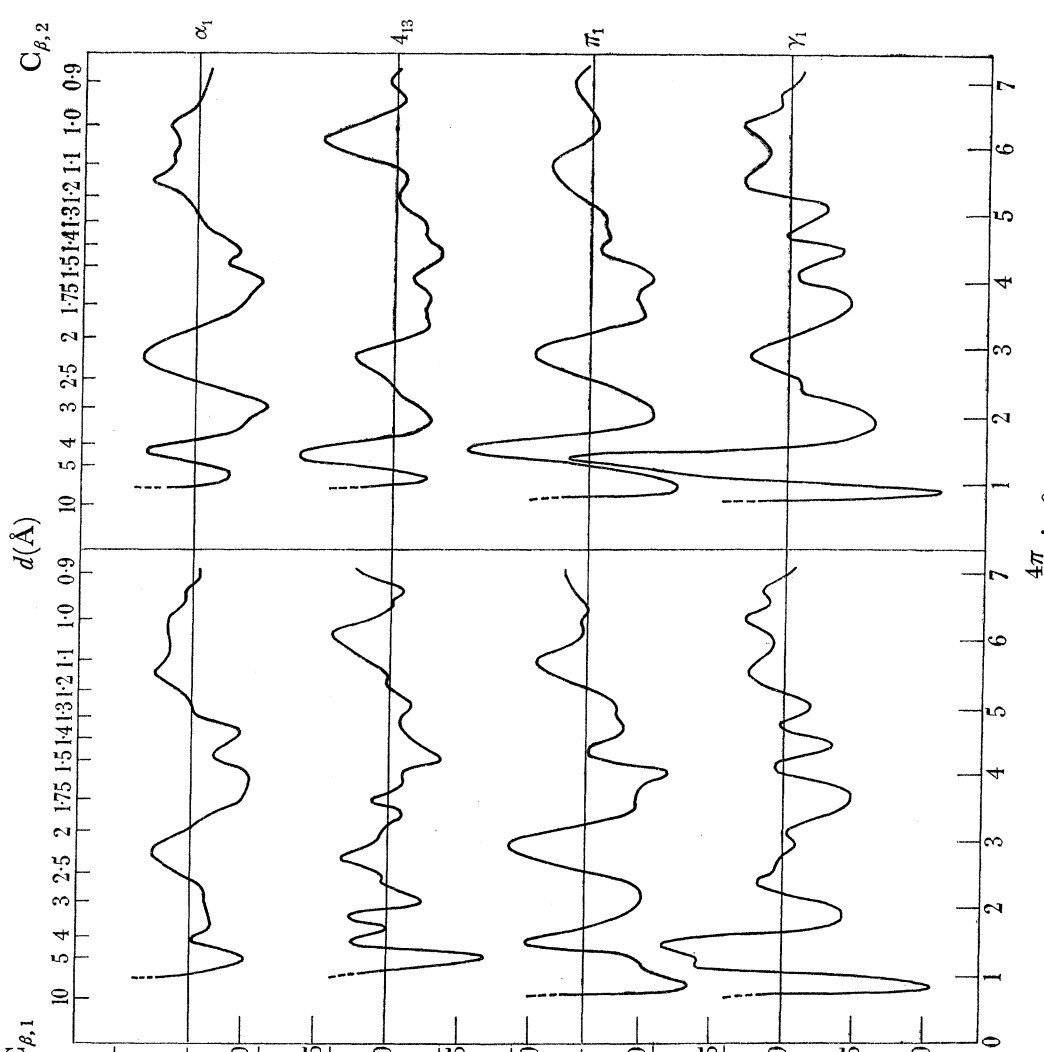

FIGURE 4. Interference intensity functions, $i(s)$, calculated for independent α, 4_{13}, π and γ helices of infinite length in random orientation. The left-hand set of curves refers to models in which the C_β-atoms are in position 1; the right-hand set refers to the alternative C_β position 2.

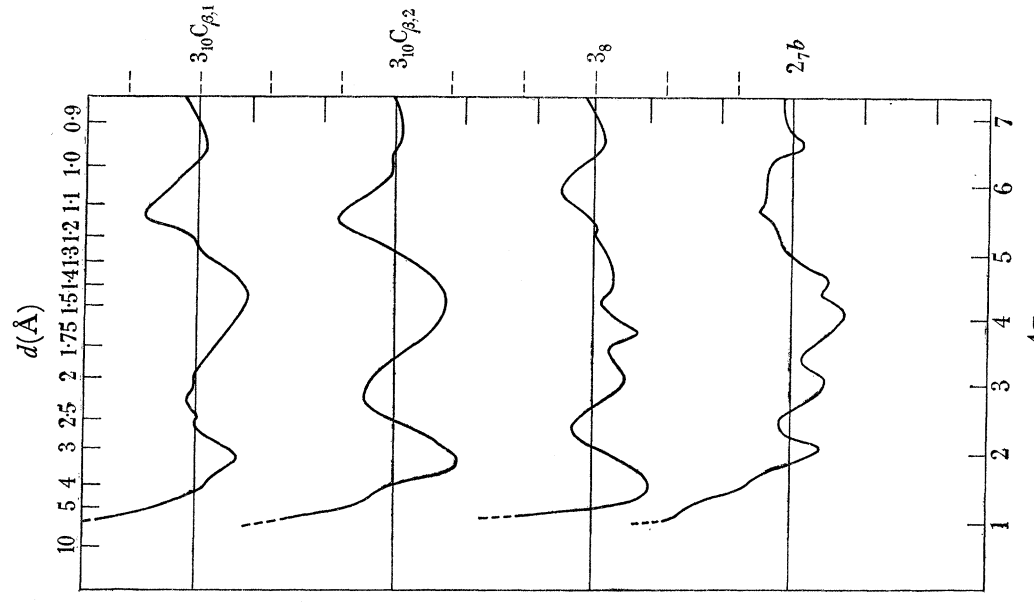

FIGURE 5. $i(s)$ functions for the 3_{10}, 3_8 and 2_7 helices; two alternative C_β positions are considered for the 3_{10} helix. The ordinate scale for all curves is as marked for figure 4.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

This connate disadvantage is lessened where regularly constructed high polymers are concerned. A linear polymer may be considered to possess interatomic distances l_{ij} and atomic number products W_{ij} of two types. The first arise from atoms within a single monomeric unit, considered to be invariant, and are evidently the same for all chain configurations; the second are due to the relative positions of different monomeric units with respect to each other. The second group of l_{ij} and W_{ij} are defined by the nature of the chain folding, and it is improbable that homometric sets could be produced by different configurations. Apart from the initial region, therefore, different radial distribution curves must result from unlike chain configurations, but the disparity may not always be striking. By virtue of the process of Fourier transformation, the differences are localized and accentuated in the corresponding $i(s)$ curves and are particularly evident in the part between $s \simeq 1$ and $s \simeq 3{\cdot}5$. It is precisely this part of the intensity curve that can be measured with the highest accuracy in normal experiments.

In the derivation of a radial distribution curve from the experimental data, equal weight is given to all parts of the intensity curve, although the precision of their measurement varies. It is consequently less easy to make allowance for such uncertainties when testing a structural hypothesis. Again, the function directly obtained from experimental measurements corresponds to $\phi(r)-\phi_0(r)$, where $\phi_0(r)$ refers to the radial distribution function corresponding to uniform intramolecular electron density. Calculation from a model, on the other hand, leads immediately to $\phi(r)$. Comparison is not thereby invalidated but it is certainly rendered less satisfactory. The equivalent discrepancy with two $i(s)$ curves is easily recognized as it is confined to the low-angle region. Finally, the computation of $i(s)$ via equation (8) is based on Fourier transformation of the convergent function $\phi(r)/r$, with a subsequent division by s. This is a sounder procedure than is involved in the converse case, which requires Fourier inversion of the abruptly terminated function $si(s)$ with the possible occurrence, as a consequence, of false features in the resulting radial distribution curve. The frequently used device of multiplying $si(s)$ by an arbitrary 'fade-off' factor before inversion can remove true as well as false features.

These arguments conduce to a preference for using $i(s)$ curves as the means of detailed comparison of theory with experiment, and it is further concluded that the method is intrinsically sufficiently sensitive to permit decision between even very similar alternative structures.

(*e*) *Contribution of the R′-atoms*

Only the C_β atoms of the side-chains have been included in the calculations in this section. The remaining R'-atoms contribute importantly to the scattering by an actual protein molecule, and it is necessary to make an assessment of their influence. This cannot be done with precision as long as only incomplete structural hypotheses are available. Certain lines of argument can, however, lead to an anticipation of the general nature of the effect.

In the corpuscular proteins commonly considered, there are encountered some twenty-one R-groups differing in chemical constitution, often to a substantial extent. Owing to the presence of single bonds, whole or partial rotation about them must be a frequent occurrence and will probably be mainly decided by stereochemical factors. These factors will vary, for a given R-group, with the nature of its immediate environment. There is

thus a substantial element of irregularity in the positions of the R'-atoms. This would be expected to be less for the simple synthetic polypeptides, where all R-groups have the same constitution, but would not be entirely absent. Irregularity in atomic positions implies diffuseness in terms of intensity of scattering, the analogy being that of a liquid or glass. In a general sense, then, one may expect that the contribution of the R'-atoms will take the form of one or two diffuse maxima and minima in $i(s)$, and it seems unlikely that these will seriously confuse the detailed pattern due to the much more regular structure of the core. A strong argument that such confusion is, in fact, not produced can be developed from the close similarity between actual scattering curves from many different α-proteins (see §6 (*c*)). The matter will be discussed further in §8 (*d*).

(*f*) *Interchain interferences*

Helically coiled chain structures presumably have a tendency to pack together as bundles of rods held together by —S—S— bridges, and Coulomb and van der Waals forces. Stereochemical factors, particularly affecting the side-chains, possibly play an important part in defining the nature of the packing. The regularly coiled polypeptide-chain cores are therefore separated by at least the average length of the R-groups, and it is impossible that all pairs of neighbouring chains possess identical mutual registrations. The result is that the inter-chain W_{ij} refer to relatively long l_{ij} which are fairly evenly distributed. To a good approximation, the interchain interferences can be considered to arise between hollow cylinders (the cores) separated by gaps of low and uniform electron density (the R'-groups). As is discussed in detail in §9, it follows that the part below $s \simeq 1$ in the $i(s)$ curves of figures 4 and 5 would, in actuality, be replaced by a region containing a definite peak, characteristic of the way in which the helices are grouped together to form a complete molecule. Such a 'packing' peak is always observed in practice.

5. Experimental procedure

(*a*) *Measurement of intensity*

The accurate measurement of the intensity scattered from an amorphous organic substance presents some difficulty, especially since strictly monochromatic radiation must be employed. At high angles, the scattering effects are very feeble but are nevertheless important (see §6 (*b*)). The low-angle region is of even greater interest as it reflects the chain-packing phenomena described in §§4 (*f*) and 9. It is therefore desirable to use a technique which permits, in one continuous experiment, the recording of intensity over an unusually wide angular range. This necessitates the use of a parallel-sided block specimen used in a symmetrical transmission setting. Of course, the intensity finally determined must refer to the protein sample only. A convenient subtractive technique for eliminating the effects of air and slit scattering is only possible using the counter method of intensity measurement, which also commends itself because of its inherent precision. In order to achieve monochromatic conditions with little loss of available intensity or of convenience, a method was developed employing a proportional counter as the X-ray detecting device.

In general, Cu $K\alpha$ radiation ($\lambda = 1{\cdot}542$ Å) was used, although a few runs were done with Ag $K\alpha$ rays ($\lambda = 0{\cdot}561$ Å) in order to extend the range of the angle variable s. The Cu

radiation was generated in a rotating-anode tube of high brilliance specially constructed for the purpose. The specimens were examined in the form of tightly compressed compacts of the powdered material or of finely chopped-up fibres; containing windows were not found to be necessary. The intensity was measured by means of a proportional counter used in conjunction with a single-channel pulse amplitude analyzer adjusted so that only quanta in a narrow band centred on $K\alpha$ were counted. When the usual $K\beta$ filter was employed, the degree of monochromatization achieved was *ca.* 99% in terms of energy (Arndt & Riley 1952). The diffractometer was fitted with a monitoring counter in the primary beam. In this way the scattering from the specimen was measured in terms of the primary intensity and was not affected by variations in the latter. Details of the diffractometer and of its use have been described elsewhere (Arndt, Coates & Riley 1953), together with an account of the correction factors to be employed. In the present work, finer slits were used than there specified, in order to ensure that the degree of angular resolution was more than sufficient. In addition, greatly improved stability was achieved by the use of the continuous-flow type of counter described by Arndt, Coates & Crathorn (1954).

(*b*) *Reliability of results*

The intensity curves of § 6 refer to the protein sample only (including, of course, moisture and impurities) and have been corrected in the usual way for experimental factors. They include the incoherent Compton scattering, but the effect of fluorescent radiation is negligible. Counting losses can be disregarded when a proportional counter is used, and the total number of counts recorded in the experiments was such that statistical inaccuracies were unimportant. The monitoring device maintained an excellent degree of constancy in the basic intensity. In short, the only appreciable source of error or variation ascribable to the apparatus that could be detected under normal operating conditions was due to occasional small drifts in the counting circuits. Checks for this effect were a matter of routine.

The degree of reproducibility of results for a given sample, if in the form of a fine powder, was very high, even when conditions were deliberately varied in order to constitute a test. The accuracy of measurement of scattered intensity, over a range of s from about 0·1 to 6·5, was within 1% in the most favourable cases, although for the purposes of rapid survey a lower accuracy was accepted. The intensity curve for serum albumin is particularly trustworthy. This protein, in the form of finely powdered air-dry crystals, was adopted as a standard substance for testing the apparatus and the various correction factors, and its intensity curve has been repeatedly determined.

All curves show, as theory predicts, a general lowering of intensity towards the higher values of s. As is discussed in § 6 (*b*), the absolute scale of intensity is determined by this high s region, and it is important that it be known accurately. Errors arise in its determination because of the almost grazing angles of incidence that are necessitated in the symmetrical transmission setting. Heavy reliance must therefore be placed in the efficiency of the geometrical volume correction factor, $\cos\theta$, in this region. It is clear that, if the specimen deviates from the assumed shape of a block with perfectly plane and parallel sides, the correction factor will not operate satisfactorily. In particular, any bowing of the specimen into a concave or convex shape will seriously affect the extreme high-angle part

of the intensity curve and a satisfactory correction procedure will be impossible. The low- and medium-angle regions, on the other hand, will not be greatly affected as the change of irradiated volume with angle is there much less marked.

The alternative reflexion setting of the specimen is satisfactory as regards the high-angle region but encounters similar difficulties in the range of moderate and low angles. The nature of the absorption correction is more complicated in this case unless a specimen of 'infinite' thickness be used, an impossibility with organic substances owing to their low absorption coefficients. In our experience, the transmission setting gives a reliable curve except at extremely high angles ($2\theta > 160°$) and has the advantage that the absorption correction is easily applied. With serum albumin, the reflexion setting was used to check the higher angle region of scattering but was soon abandoned as it provided no real gain. The exploration of the general trend of the scattering curves at high s values is best done by using a shorter wave-length. The scattering for serum albumin was measured with Ag$K\alpha$ radiation up to $s \simeq 13$; although not very accurate, the results were quite adequate to confirm the general level of the high-angle region between $s \simeq 5$ and $s \simeq 6{\cdot}5$, as determined with Cu$K\alpha$ radiation.

There is one other source of error against which it is difficult to guard. It is impossible to be certain that a block specimen of dimensions *ca.* $10 \times 10 \times 1$ mm, prepared by compression of a powder, is completely homogeneous in density. A linear absorption factor measured over the narrow area of the beam at $2\theta = 0°$ may, therefore, not be representative of the whole specimen as it rotates. The effect can be examined by measuring the absorption at a number of angular positions of the specimen relative to the primary beam, so that different paths through it are involved. With compacts made from finely powdered air-dry crystals, the variation in absorption with angle was small and irregular and the adoption of an average value was permissible. Flaky specimens, which are often produced from freeze-dried samples, gave a less satisfactory result. The effect of adopting a wrong absorption correction factor is progressively to raise or lower the corrected intensity curve with increase of s. Without repeating each run a number of times with specimens of different thicknesses, as was done with serum albumin, it is difficult to be sure that this has not happened to some extent with a typical sample.

In summing up, it may be said that the measurement of the *whole* course of the corrected intensity curve is necessarily less accurate than that of a limited region. Even slight singularities in the curve can be established beyond doubt, but the exact form of the monotonic curve on which they are superimposed is not capable of determination with equivalent precision. Even so, in the carefully measured curves which are used for illustration in this paper, the discrepancy is unlikely to exceed $\pm 5\%$ in the normalized 'absolute' height of the principal peak at $s \simeq 1{\cdot}5$ (see § 6).

6. Intensity curves for various proteins

(*a*) *The three principal types*

A survey of a number of proteins shows that the scattering curves from isotropic specimens fall into only three essentially different types which we propose to call α, β and γ. Figure 6 shows examples of the three types drawn, within practical limitations, on the

same normalized scale of intensity. The differences between the curves are considerable and lie far beyond the bounds of any kind of experimental error. A cursory inspection of its scattering curve is sufficient to assign a protein to its appropriate class and, in particular, variations within the α class appear to be minor.

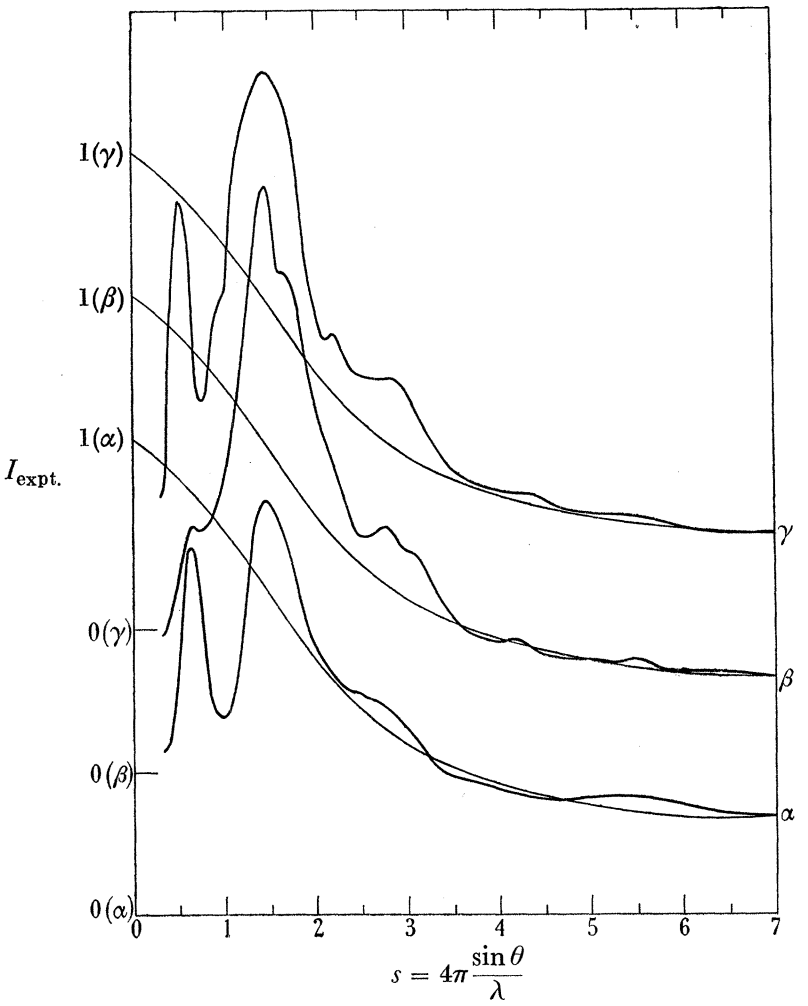

FIGURE 6. Comparison of intensity curves, $I_{expt.}$, given by examples of the three principal types of proteins: α, serum albumin; β, silk fibroin; γ, collagen. The smooth thinner curves represent $I_{tot.}$, the coherent plus incoherent scattering from independent atoms. Zeros displaced as indicated.

The initial portion of the curves, notably the peak at $s \simeq 0{\cdot}5$, is not the concern of this section. In the α-proteins, it arises from the inter-chain interferences discussed later (§9) and to which reference has already been made in §4(f). It is the remaining part of each curve, from $s \simeq 0{\cdot}75$ outwards, which is characteristic of the structural class and to which attention is drawn here.

All types exhibit an important peak at $s \simeq 1{\cdot}5$. This principal peak in an α-curve is usually considerably lower in height than in the other types; it is completely smooth and possesses a definite asymmetry, its side of steeper ascent lying towards low s values. By contrast, the equivalent peak for β-proteins is always complex and possesses two, and

sometimes three, components. With silk fibroin, as illustrated in figure 6, a pronounced shoulder at $s \simeq 1{\cdot}65$ is clearly visible. The γ-type curve shown by collagen exhibits two flanking subsidiary peaks, at $s \simeq 0{\cdot}85$ and $s \simeq 2{\cdot}1$, on either side of the main peak. A β-curve is generally richer in definite peaks than an α-curve; the far-out peak at $s = 5{\cdot}45$ (or 1·15 Å), shown by silk fibroin, is a typical β-characteristic never observed with α-proteins. The structure of β-proteins will be the subject of a future paper, but it may be said here that there is little doubt, both from our own investigations and those of others, that the basis of such structures is the pleated-sheet hypothesis of Pauling & Corey (1951 *b*, 1953).

Table 2 classifies into their appropriate structural types the proteins examined in this work. Astbury's (1940, 1953) tripartite scheme of classification, based principally on examination of fibrous proteins, thus seems to be generally applicable. The overwhelming majority of corpuscular proteins are of the type called *k-m-e-f* (keratin-myosin-epidermin-fibrinogen) by Astbury. (N.B. In the alcohol-dried form examined by us, myosin definitely gave a β-type scattering curve.) Results on the *A* and *B* fractions of insulin (Sanger 1949) are also included, and it will be noticed that a change of configuration has occurred during their extraction from the complete insulin molecule.

TABLE 2. STRUCTURAL CLASSIFICATION OF PROTEINS EXAMINED

α	
chymotrypsinogen (ox)	insulin (ox)
clupein (herring sperm)	keratin (α-form)
cytochrome-*c* (horse)	β-lactoglobulin
edestin	lysozyme
egg albumin	myoglobin (horse, whale)
fibrinogen (man)	pepsin
γ-globulin (man, ox)	ribonuclease
haemoglobin (man, horse, ox)	serum albumin (ox)
histone (calf-thymus)	
β	
actin	insulin-*B*
α-casein	myosin
insulin-*A*	silk fibroin
γ	
collagen (rat-tail, oxhide)	gelatin (pigskin)

(*b*) *Determination of absolute scale of intensity*

In order to place an intensity curve on an absolute scale it is necessary to know, at least approximately, the elementary composition of the specimen. The intensity of coherent plus incoherent scattering from hypothetical material of this composition, in which all the atoms are presumed to scatter independently, is first computed from tabulated atomic scattering factors of both types, and is called $I_{\text{tot.}}$. At high values of s, the observed curve $I_{\text{expt.}}$ will approximate to this total independent scattering curve because of the unimportance of interatomic interference effects in this region. The procedure is to match $I_{\text{expt.}}$ with $I_{\text{tot.}}$ over a range of high values of s and thus place the two curves on the same scale. It is convenient to express $I_{\text{tot.}}$ in normalized form such that its value is unity at $s = 0$; in this way, different experimental curves can be easily compared absolutely on a normalized intensity scale.

Proteins vary little in elementary composition, and it can be shown that the range of differences that exist has a negligible influence on $I_{\text{tot.}}$. For our standard curve we

adopted Osborne's (1902) analyses for horse-serum albumin adjusted to a water content of 10%. The empirical formula of the hydrated molecule is then $C_{0\cdot60}H_{1\cdot00}O_{0\cdot21}N_{0\cdot15}S_{0\cdot01}$ as referred to unit value for H. This standard $I_{\text{tot.}}$ curve is drawn lightly in figure 6 for each of the three types of $I_{\text{expt.}}$ illustrated.

Once $I_{\text{expt.}}$ has been normalized in this way, $i(s)$ can be derived from the relation

$$i(s) = \frac{I_{\text{expt.}} - I_{\text{tot.}}}{I_{\text{coh.}}}, \tag{11}$$

where $I_{\text{coh.}}$ represents the coherent part of the normalized independent scattering. It is evident that $I_{\text{coh.}}$ is equivalent to $A(s)$ (equation (3)), and the numerator in (11) to $B(s)$.

The principal disadvantage of the procedure is the subjective nature of the process of curve-matching. The greater the range of s embraced in $I_{\text{expt.}}$, the better the matching is likely to be, but overmuch effort spent on always extending the s range would be wasted owing to the limited accuracy with which atomic scattering factors are known. The curves obtained with Ag $K\alpha$ radiation extended out to $s \simeq 13$ and allowed a reasonably satisfactory matching procedure to be established for use with all the usual Cu $K\alpha$ curves of lesser range. In all cases, the $I_{\text{expt.}}$ and $I_{\text{tot.}}$ curves were taken to be coincident in the neighbourhood of $s \simeq 6$.

Uncertainty is obviously involved in the derivation of $i(s)$ from an experimental intensity curve, both because of the matching procedure and by virtue of the problems discussed in §5(*b*). The uncertainty, however, only affects the position of the whole curve relative to the zero abscissa; it lifts or lowers the curve roughly as if it were hinged about the matching point. The number of singularities in the curve remains the same, as does their shape; only their relative importance can be modified.

(*c*) *α-Proteins*

The intensity curves given by all the proteins called α in table 2 are very similar and, for the purposes of detailed discussion, we have selected six only, from serum albumin (ox), insulin (ox), haemoglobin (ox), myoglobin (horse), myoglobin (whale) and ribonuclease. The intensity data were determined with particular care, each protein being examined several times. The normally accepted molecular weights vary from about 6000 (or, perhaps, 12000) to 70000, and one protein, myoglobin, is present in two varieties from different species. Each scattering curve was first normalized in the usual way, independently of the others, and the results are depicted in figure 7. Discussion of the inter-chain interferences producing the inner peak is deferred until §9.

It will be observed that the curves are all exceedingly alike and differ principally in the normalized height of the inner half of the curve and, in particular, of the peak at $s \simeq 1\cdot5$. This is precisely the kind of variation that is caused by the factors discussed in §5(*b*), and it is consequently difficult to be certain that such divergences are real or, if real, are exactly of the indicated magnitude. After a careful appraisal of the factors involved and a minute assessment of the data for each protein, we have come to the conclusion that divergences of this type do definitely exist but that it is impossible, without much further work, to give them precise numerical values. Be that as it may, it is evident that the similarities between the curves are of greater importance than any conceivable differences.

The more subtle features of the scattering curves are made more apparent in the mean interference intensity function $i(s)$ reproduced in figure 8. This curve was obtained by averaging the six $i(s)$ curves corresponding to figure 7. By this process, any interference effects due to the common predominant polypeptide chain configuration would be expected to be enhanced at the expense of any specific side-chain effects. The important

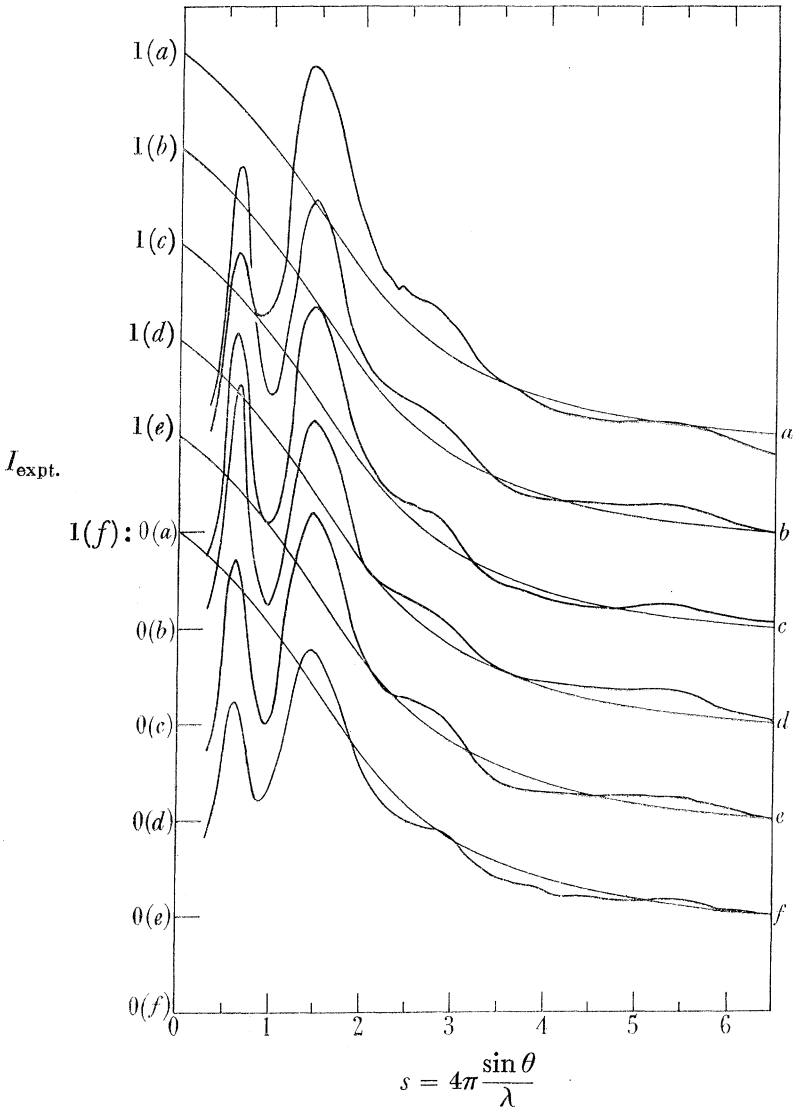

FIGURE 7. Intensity curves, $I_{\text{expt.}}$, given by six different α-proteins: *a*, ribonuclease; *b*, haemoglobin (ox); *c*, serum albumin (ox); *d*, myoglobin (whale); *e*, myoglobin (horse); *f*, insulin (ox). N.B. The peak at $s = 0{\cdot}6$ for ribonuclease lies below the much higher peak for ox haemoglobin and should not be confused with it. Zeros displaced as indicated.

features are: first, the three principal maxima, *A*, *B* and *C*, of which the last is seen to possess a central depression in most of the individual curves, e.g. that for serum albumin, also reproduced in figure 8; the next important feature is the slight, but definite, convexity *D*; thirdly, some of the curves, but not all, exhibit the slight step at *E*; finally, all show the shoulder *F*. The subsidiary convexities *G* and *H* are sometimes present.

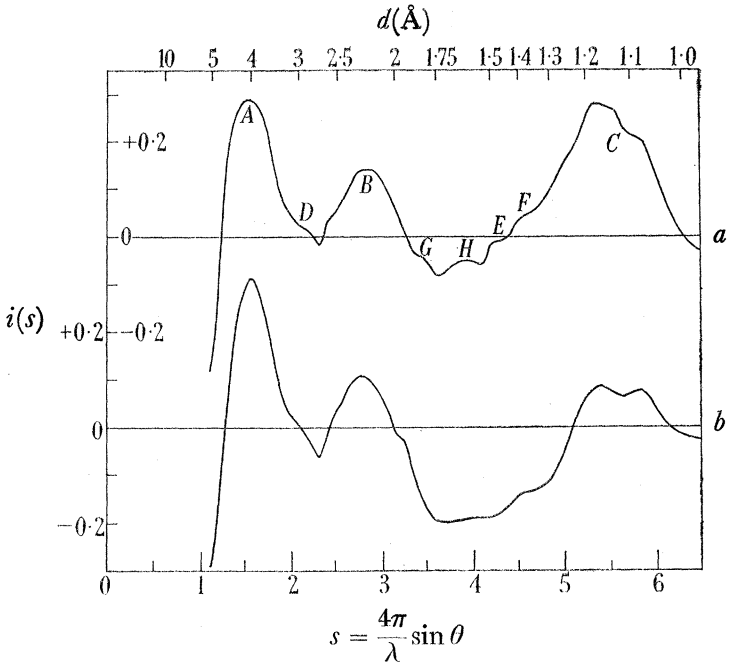

FIGURE 8. Interference intensity functions, $i(s)$ for α-proteins. (a) is a composite curve obtained by averaging the six $i(s)$ functions derived from the $I_{\text{expt.}}$ curves in figure 7. (b) refers to serum albumin alone.

7. GENERAL CONSIDERATIONS IN THE COMPARISON OF $i(s)$ CURVES

The significance of the definition (11) of $i(s)$ derived from $I_{\text{expt.}}$ (§6(b)) needs to be examined to make clear its relation to the theoretical $i(s)$ functions of §4(c) calculated from molecular models. Only the C_α, C_β, C′, N and O atoms of the so-called core of the polypeptide chain are involved in the derivation of $i(s)_{\text{theor.}}$ whereas all the atoms present in the specimen are concerned in defining $i(s)_{\text{expt.}}$. For the sake of simplicity, it will be assumed that the specimen contains only protein and water, and has the atomic composition already mentioned in §6(b). Schematically, the scattering from the specimen may be divided into two parts: one due to core atoms (and pairs of atoms) equivalent to those dealt with in a structural hypothesis, and the other to all the other atoms (and pairs of atoms) actually present. The coherent part of the experimental intensity may therefore be written as

$$_{\text{coh.}}I(s)_{\text{expt.}} = A(s)_{\text{core}} + A(s)_{\text{rem.}} + B(s)_{\text{core}} + B(s)_{\text{rem.}}, \tag{12}$$

where the subscript rem. refers to the remainder of the atoms in the hydrated molecule after exclusion of the C, N and O core-atoms. The H atoms of the core are included in this remainder together with all the atoms in the side-chains other than C_β, as well as the water atoms. $B(s)_{\text{rem.}}$ includes, of course, cross-terms between 'core' and 'remainder' atoms.

The procedure adopted in the derivation of $i(s)_{\text{expt.}}$ according to (11) is equivalent to writing

$$i(s)_{\text{expt.}} = \frac{B(s)_{\text{core}} + B(s)_{\text{rem.}}}{A(s)_{\text{core}} + A(s)_{\text{rem.}}}, \tag{13}$$

since the incoherent scattering cancels out. The function with which the model calculations of §4(c) are concerned is, evidently,

$$i(s)_{\text{theor.}} = \frac{B(s)_{\text{core}}}{A(s)_{\text{core}}}. \tag{14}$$

The interference intensity function due to the remainder atoms (considered as part of the complete structure, and not in isolation) may be defined by

$$i(s)_{\text{rem.}} = \frac{B(s)_{\text{rem.}}}{A(s)_{\text{core}} + A(s)_{\text{rem.}}}. \tag{15}$$

If, therefore, the simplification adopted in § 3 (*a*) is again used, whereby all the atomic scattering are factors assumed to have the same form (see equation (6)), we may write

$$i(s)_{\text{expt.}} = \frac{\nu_{\text{core}}}{\nu_{\text{tot.}}} i(s)_{\text{theor.}} + i(s)_{\text{rem.}}, \tag{16}$$

where ν_{core} represents the sum of the squares of the atomic numbers of the C, N and O core atoms only and $\nu_{\text{tot.}}$ refers equivalently to all the atoms present in the hydrated protein specimen. For brevity, the ratio $\nu_{\text{core}}/\nu_{\text{tot.}}$ will be called μ.

An $i(s)$ curve obtained by experiment can therefore be regarded as the sum of two component curves, one of which is exactly equivalent to a scaled-down version of a theoretical function calculated from a model while the other is produced by all the terms omitted from such calculations. If the molecular model adopted is near the truth, the latter function is, in principle, given by

$$i(s)_{\text{rem.}} = i(s)_{\text{expt.}} - \mu i(s)_{\text{theor.}}. \tag{17}$$

In practice, the derivation of an exact difference function is made difficult because of the uncertainties arising in $i(s)_{\text{expt.}}$ that were discussed in §§ 5 (*b*) and 6 (*b*).

The factor μ is capable of reasonably precise evaluation for the α-proteins if it is assumed that the whole of each polypeptide chain is folded into the particular configuration for which $i(s)_{\text{theor.}}$ is calculated. Its value is then close to 2/3, but it is clear that a lower figure will result if some parts of the core of the polypeptide chain possess a different configuration from the main portions, as would, for instance, arise from the presence of sharp bends in the folded chain. In such a case, the scattering from the bends must be included in $i(s)_{\text{rem.}}$.

8. Polypeptide chain configurations in α-proteins and polypeptides

(*a*) *Initial examination of $i(s)$ curves for α-proteins*

It may be expected, from the considerations advanced in § 7, that success in distinguishing within an actual $i(s)$ curve the features essentially due to the chain configuration will depend as much on the nature of the scattering from the remainder atoms as of that from the core. In general, the features sought will be superimposed on a background constituted of $i(s)_{\text{rem.}}$ and may, in part, be confused with it. If, however, $i(s)_{\text{core}}$ possesses pronounced characteristics in a region where $i(s)_{\text{rem.}}$ is lacking in detail or is unimportant, there will be no difficulty in recognizing them. As regards the synthetic polypeptides, the indications are sufficiently clear-cut (figure 10) over the whole range of s that immediate choice of a single model is strongly suggested. It is permissible, from the genetical resemblances between the scattering curves, to consider the polypeptides as being 'type specimens' of the more complicated protein structures, and to take advantage of this fact in elucidating the significance of the less clear protein results. While this will be done later, it is desirable

first to consider the data for the proteins in isolation in order to see how far an independent approach will lead.

The first important conclusion to be drawn from the close similarity between the intensity curves for all α-proteins is that the same fundamental polypeptide chain configuration is almost certainly present as the predominant component in every case. It is reasonable to suppose that the nature of this predominant configuration should be revealed by comparing the observed curves with those calculated for various structural models.

By far the most important property of the experimental intensity curves for the α-proteins is the maximum A at $s \simeq 1{\cdot}5$ (figure 8), which we shall call the 'principal peak'. This typical and pronounced peak must be primarily a function of the intrinsic regularity of folding of the polypeptide chain. Although it may be affected by the scattering due to the remainder atoms, the principal peak is certainly the central criterion to be used in distinguishing between the various configurational hypotheses. The fact that it is absent in the curves of figure 5 allows the immediate exclusion of the 3_{10}, 3_8 and 2_7 helices from further consideration. There are, it is true, slight hesitations in the upward sweep of the low-angle intensity in the two 3_{10} curves, but the effect is altogether too trivial to give rise to the actually observed peak under any circumstances.

No other structure can be excluded on this simple basis, but a further rapid examination of the various $i(s)$ curves is sufficient to rule out the $4_{13}(C_{\beta,1})$ and γ_2 helices (figure 4). In the former case, a marked extra peak at $s \simeq 1{\cdot}9$ accompanies that at $s \simeq 1{\cdot}5$ and would certainly be apparent in the experimental curves; its complete absence in actuality constitutes ample grounds for rejecting the $C_{\beta,1}$ variant of the 4_{13}-helix. The principal peak in the case of γ_2 is doubled by the addition of a shoulder at $s \simeq 1{\cdot}1$ which is not observed in practice. An additional reason for discarding the γ_2-helix is that the next broad maximum contains three peaks at $s \simeq 2{\cdot}4$, $2{\cdot}8$ and $3{\cdot}2$, of which the first is the most pronounced. This region is quite dissimilar to the single broad peak B centred on $s \simeq 2{\cdot}8$ in the experimental curves.

Reasoning based on the normalized height of the principal peak is less convincing, as it must obviously not neglect to consider the following points:

(a) the operation of the μ factor in diminishing the scale height, but not the relative height, of peaks in $i(s)_{\text{theor.}}$;

(b) the intervention of $i(s)_{\text{rem.}}$;

(c) the uncertainty regarding the scale height of the peak in $i(s)_{\text{expt.}}$, because of experimental and technical factors (§5(b)).

It is nevertheless clear that none of these factors could reasonably be expected to reduce the extreme height of 1·55 for the principal peak in the case of the γ_1-helix to a value of about 0·3 as reliably observed with serum albumin and several other α-proteins. While the corresponding heights, 0·89 and 0·65, are still considerable for the π_1 and $4_{13}(C_{\beta,2})$ helices, they are not so great as to compel rejection on this ground alone. Nor does the exceptionally low height of the equivalent α_2 peak necessarily exclude this structure, although it is improbable that it could be the sole configuration present.

To sum up the results of these comparisons based on the more obvious features of the $i(s)$ curves, the following configurations are left for further examination: α_1, α_2, π_1, π_2, $4_{13}(C_{\beta,2})$. The rest can be definitely rejected.

(*b*) *Comparison of selected i(s) curves*

In order to proceed to a choice between the remaining alternatives, regard must be had to the more subtle characteristics of the interference intensity functions.

Perutz (1951) has drawn attention to the importance of a crystal-type reflexion of spacing 1·5 Å from oriented chains in characterizing the α-helix. This spacing corresponds to the axial distance separating neighbouring amino-acid residues in their regular succession along the chain. The equivalent spacing for the 4_{13}-helix is at 1·4 Å and, for the version of the π-helix considered by us, at 1·14 Å. In patterns from randomly oriented chains, the criterion has less force as the peaks in question do not necessarily appear as such in the appropriate $i(s)$ curves. Nevertheless, the $i(s)$ curves for the α-helices do include a peak at 1·5 Å, or slightly less ($s = 4{\cdot}25$) because of its step-wise character. A peak of this description is certainly not a striking and general feature of the experimental curves, but a step at this position is exhibited by the two myoglobins and by ox haemoglobin, and is clearly visible (E) in the composite curve in figure 8.

Another distinction between the α_1 curve and the others is afforded by the region between $s \simeq 1{\cdot}7$ and 2·3. In the $i(s)$ function for α_1, a secondary convexity occurs before the curve descends to form a rather sharply pointed minimum, in contrast with the smooth and rounded shape of this first dip in the π and $4_{13}(C_{\beta,2})$ curves. This feature is definitely observed experimentally and is clearly seen in the curve in figure 8 for serum albumin but is shifted slightly to higher s values. The characteristic dip in the final broad maximum in the α-helix curves is also commonly found in the α-protein curves.

It is our opinion, based on the examination of numerous intensity curves for corpuscular proteins, that the configurational model in best agreement with the experimental data is the α_1-helix. All of its features are observed in one or other of the α-proteins examined and no gross discrepancies are apparent. This conclusion is reinforced if one takes account of the data for the synthetic α-polypeptides. Nevertheless, on the basis of the protein data alone, it is impossible to assert that the π_1, π_2 and $4_{13}(C_{\beta,2})$ helices are markedly inconsistent with observation. Equally, the admixture of a certain proportion of α_2 would not be an impossibility, although a 100 % α_2 structure would seem to be highly improbable.

The length of an undeviated coiled chain in a corpuscular protein can never be very great because of the limited size of the molecule. We have calculated, but not reproduced, $i(s)$ for a short α-helix containing only twelve residues. The resultant curves are very similar to those illustrated for a helix of infinite length, but all the features are somewhat more blurred. An important difference is that the principal peak is appreciably broader in the new α_1 curve and thereby more in conformity with that observed in practice.

In figure 9 the curve for an α_1-helix is compared with the composite curve already shown in figure 8. Figure 9 should be studied in conjunction with figure 10 which refers to the synthetic α-polypeptide, poly-γ-benzyl-L-glutamate.

(*c*) *Synthetic α-polypeptides*

Even a cursory examination of the $i(s)$ function given by poly-γ-benzyl-L-glutamate (α-form) is sufficient for its strong qualitative resemblance to the curve for an α_1-helix of infinite length to be perceived. Figure 10 shows that each feature in the theoretical curve is matched by an equivalent one in the experimentally observed function, the important

difference being in the height of the principal peak at $s \simeq 1{\cdot}5$, or 4·2 Å in 'spacing' terms. This particular discrepancy is the only marked one and may be attributed to the relatively greater importance of the contributions from the 'remainder' atoms in this synthetic polypeptide with its large side-chains. It should be noted that the peak at 1·5 Å is

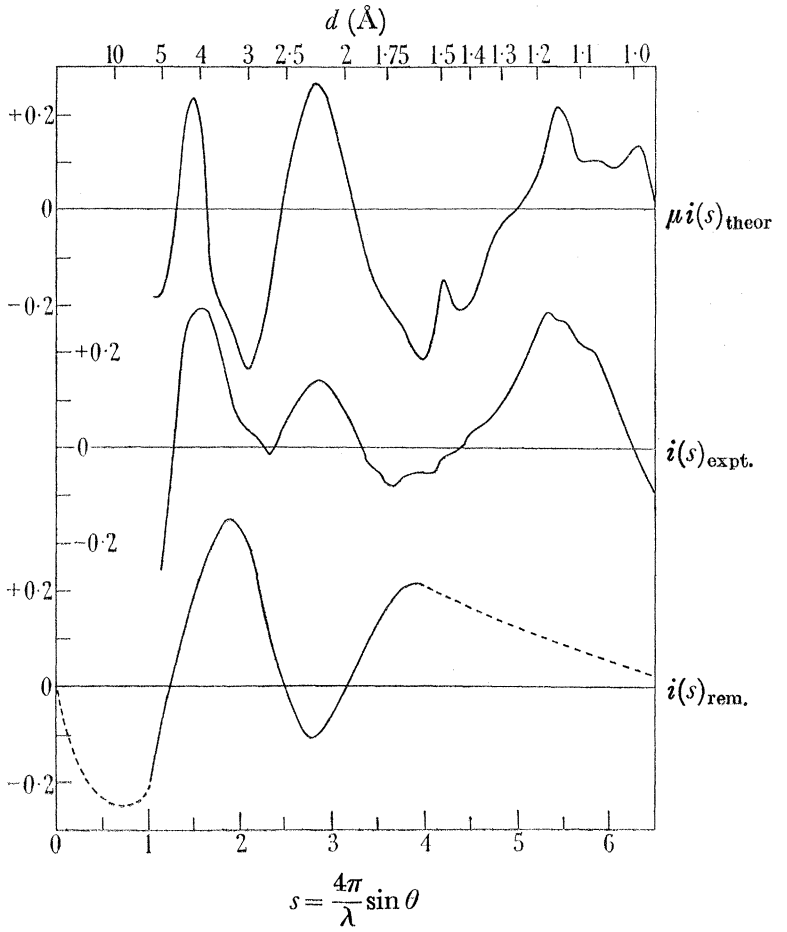

FIGURE 9. Comparison of theoretical and experimental $i(s)$ curves for α-proteins, where $\mu i(s)_{\text{theor.}}$ refers to an α_1-helix and is appropriately scaled, and $i(s)_{\text{expt.}}$ is the composite curve of figure 8. $i(s)_{\text{rem.}}$ represents the effect of the 'remainder' atoms not included in the calculations, i.e. principally the non-C_β atoms in the side-chains.

prominent in this case. This is not so with the very similar compound poly-γ-methyl-L-glutamate (α-form) whose curve is not shown and which, apart from this difference, is very closely of the same character. It therefore appears that the occurrence of the 1·5 Å peak is, as in the case of the proteins, a somewhat adventitious event, possibly related to specific effects due to the side-chains.

(*d*) *Effect of the remainder atoms*

Subtraction of $\mu i(s)_{\text{theor.}}$ from $i(s)_{\text{expt.}}$ should, as stated in equation (17) of §7, lead to $i(s)_{\text{rem.}}$. The approximate value of μ for the proteins is 0·66, while for the two polypeptides it is lower: 0·47 for the methyl and 0·32 for the benzyl derivative. In figures 9 and 10 the effect of these factors has been included in the theoretical $i(s)$ plots. When the subtractions are carried out, somewhat jumpy curves result, smoothed versions of which

are shown in the lowest sections of figures 9 and 10. The precision with which the shape of these difference curves is known is naturally small, since the heights of the observed peaks in $i(s)_{\text{expt.}}$, particularly at high values of s, are sensitive to the matching procedure (§6(*b*)) as well as to the experimental factors surveyed in §5(*b*). Moreover, the low-angle part of $i(s)_{\text{theor.}}$ has no correspondence with reality, as it must be modified in the region below $s \simeq 1$ by the inter-helix packing effects described in §9. The course of $i(s)_{\text{rem.}}$ can thus only be surmised in the parts of the curves which are shown by broken lines in figures 9 and 10. It is, nevertheless, possible to derive an approximate radial distribution function from each difference curve and to make certain deductions from it.

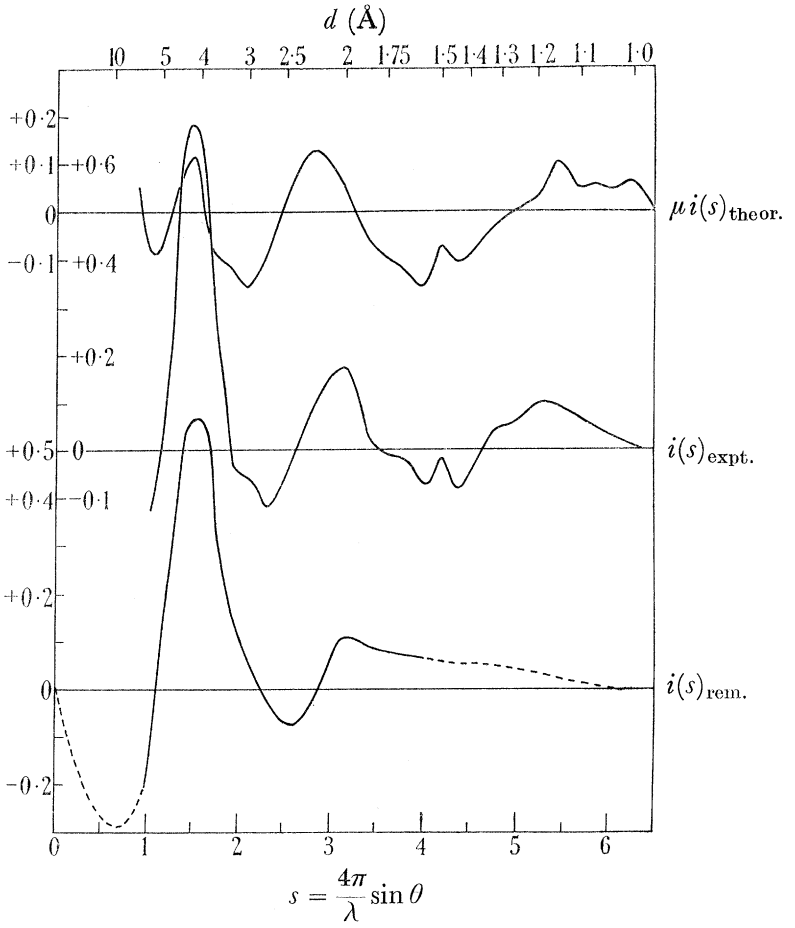

FIGURE 10. Comparison of theoretical and experimental $i(s)$ curves for poly-γ-benzyl-L-glutamate(α-form) in a similar way to figure 9.

The function obtained by Fourier transformation from $i(s)_{\text{rem.}}$ in the usual way is $[\phi(r)-\phi_0(r)]_{\text{rem.}}$ in the sense already mentioned in §4(*d*); for convenience, the function will be written as $\Delta\phi_{\text{rem.}}(r)$. As will be seen from figure 11, the main features of this radial distribution function in the case of the α-proteins are two peaks at $r \simeq 1{\cdot}6$ Å and $r \simeq 3{\cdot}9$ Å, the second of which is much more pronounced. The equivalent peaks for the synthetic polypeptide are at $r \simeq 1{\cdot}7$ Å and $r \simeq 5{\cdot}3$ Å; in this case the latter peak is much broader and unsymmetrical and is again much the more important. It is unlikely that the most prominent peak in $\Delta\phi_{\text{rem.}}(r)$ is seriously affected by the uncertainties in $i(s)_{\text{rem.}}$; the exact

position of the first peak and also detail at the higher r values are probably ill-defined, but there is no reason to doubt the reliability of the general appearance of the curve. The physical meaning of peaks in such a curve is that the corresponding interatomic distances occur at more-than-average frequency; in the terminology of § 3, the $W_{ij}(l_{ij})$ show a strong tendency to be grouped around values of l corresponding to the peak values in r. It is clear why this should be so in $\Delta\phi_{\text{rem.}}(r)$. The first peak, at *ca.* 1·6 Å, refers to the inner structure of each side-chain considered in isolation and is mainly produced by the average chemical bond distance within them; a similar peak is shown by all organic structures. The second, and most prominent, peak must relate to the average distance of side-chain

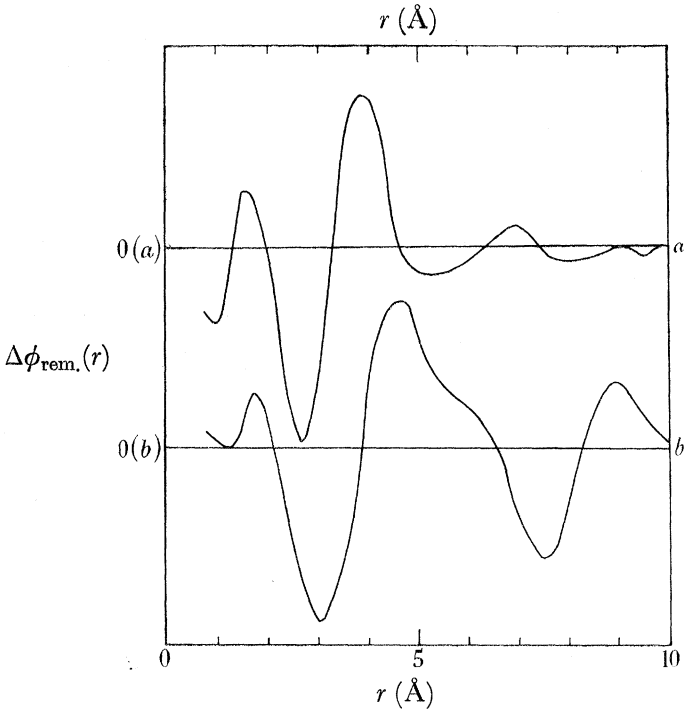

FIGURE 11. Radial distribution curves derived from $i(s)_{\text{rem.}}$ for (*a*) α-proteins, (*b*) poly-γ-benzyl-L-glutamate (α-form), showing principally intra-side-chain and side-chain-to-core-of-helix distances.

R'-atoms (i.e. excluding C_β) from the atoms of the core (including C_β). That the distance observed for the synthetic polypeptide is appreciably greater than that for the proteins is in accord with the fact that the side-chains in poly-γ-benzyl-L-glutamate are bulkier than in the proteins. Its breadth and asymmetry are also suggestive of a complicated side-chain. It would be inadvisable to reason too closely on the actual values given here, but they appear to be lower than would be expected if all the side-chains stood out at maximum extension from the body of the helix. A tendency to compactness is, in general, probable on energetic grounds, and the figures given in § 9 for the distances between neighbouring helices also support this conclusion. It should, of course, be borne in mind that $i(s)_{\text{rem.}}$ is not produced by side-chain atoms alone but by all the non-core atoms actually present.

In comparing the theoretical and experimental $i(s)$ curves it should be noted that the observed peaks, particularly the principal one, are generally broader than calculated, the effect being more marked for the proteins. This is what causes the jumps in the actual

difference curves before smoothing. It would seem that the perfection and the undeviated length of the α-helix are not as great as had been assumed in the calculations and that rather more blurred curves would give a better match with experiment. If this were so, the genuine difference curves would correspond closely to the smooth $i(s)_{\text{rcm.}}$ curves of figures 9 and 10.

9. Helix packing in α-proteins and polypeptides

The existence of a peak at $s \simeq 0{\cdot}5$ in intensity curves of the α-type (figure 7) necessarily means that the individual helices (or parts of the same helix if bent back on itself) are packed together in a more or less parallel fashion within the molecule. A single isolated α-helix does not give a peak in this region, and it can only arise from inter-helix interferences. The position of this 'packing' peak provides a measure of the average distance separating neighbouring helices, while its height and breadth can be affected by several factors, of which the most important are (*a*) degree of parallelism of adjacent helices, (*b*) number and grouping of helices in the bundle, (*c*) distance of separation of neighbouring helices compared with their diameter, i.e. the gap.

As was mentioned in §4(*f*), there can be no uniform registration between pairs of α-helices in a compact assembly. It follows that each helix can be reasonably approximated by a thin cylindrical shell and the intensity of scattering from the assembly calculated by means of appropriate relations derived by Oster & Riley (1952).

The general expression for the normalized intensity of scattering from independent aggregates in random orientation, each consisting of n identical cylindrically symmetric units of infinite length, is

$$I_{\text{norm.}}(s) = \frac{1}{n^2} f^2_{\text{cyl.}}(sR) \sum_p \sum_q J_0(s l_{pq}), \tag{18}$$

where l_{pq} represents the distance apart of the axes of the pth and qth cylinders. The double summation embraces all pairs of cylinders in the aggregate or molecule in precisely the same way as is involved in the analogous equation (1) of §3(*a*). If an α-helix be approximated by a thin cylindrical shell of radius $R = 2$ Å, the scattering factor for isolated cylinders, $f_{\text{cyl.}}(sR)$, becomes $J_0(2s)$ when l is also expressed in ångströms. Equation (18) can be used to derive the course of the inter-helix scattering curve for any postulated packing of parallel helices.

The effect of a skew arrangement is very difficult to express mathematically. If the degree of non-parallelism is not great, equation (18) will still hold approximately; at the other extreme, substantial skewness will tend to eliminate distinct packing effects by broadening and weakening the packing peak. The simple fact that this peak is usually an important feature of the intensity curves indicates that neighbouring helices are, in the main, arranged in a nearly parallel way within most protein molecules.

The experimental data concerning the position and height of the packing peak for seventeen proteins and two synthetic polypeptides are given in table 3. The peak position is, for convenience, expressed in ångströms as a 'spacing', d, derived using Bragg's law, but this familiar practice must not, of course, be taken as implying the veracity of this law as applied to the interior of a protein molecule. The most significant way of expressing the

height of the inter-helix peak is with reference to the magnitude of the intra-helix scattering; in all cases, the height of the packing peak is quoted as a proportion of that of the principal intra-helix peak at $s \simeq 1{\cdot}5$ (*ca.* 4·5 Å).

TABLE 3. THE INTER-HELIX PACKING PEAK

protein substance	apparent Bragg spacing d of packing peak (Å)	height of packing peak relative to principal intra-helix peak
cytochrome-*c*	$9{\cdot}5_3$	0·81
haemoglobin (man)	$9{\cdot}5_3$	1·00
ribonuclease	$9{\cdot}6_1$	0·63
γ-globulin (man)	$9{\cdot}6_1$	0·66
γ-globulin (ox)	$9{\cdot}6_1$	0·63
lysozyme	$9{\cdot}6_8$	0·70
serum albumin (ox)	$9{\cdot}8_4$	0·97
chymotrypsinogen	$9{\cdot}8_4$	0·77
fibrinogen (man)	$9{\cdot}8_4$	0·72
haemoglobin (ox)	$9{\cdot}8_4$	1·06
haemoglobin (horse)	$9{\cdot}8_4$	1·11
myoglobin (horse)	$9{\cdot}8_4$	1·08
insulin	$10{\cdot}0_0$	0·89
egg albumin	$10{\cdot}0_8$	0·80
β-lactoglobulin	$10{\cdot}9_1$	0·84
pepsin	$11{\cdot}5_3$	0·69
edestin	$9{\cdot}1_0$ / $10{\cdot}4_2$	0·64
poly-γ-methyl-L-glutamate (α)	$11{\cdot}7_8$	1·15
poly-γ-benzyl-L-glutamate (α)	$15{\cdot}0_0$	0·83

There is no correlation between the height and position of the packing peak. On theoretical grounds, such a correlation would be expected only if the number and type of arrangement of helices were the same in all cases and the only variation were in the inter-helix gap, as is evidently not the case. The comparatively high values of the peak height for the three haemoglobins, and for myoglobin and serum albumin, may be taken as indicating a high degree of helix-parallelism in these proteins, while the low values for ribonuclease, edestin and the γ-globulins suggest a definite skewness. In the example chosen below (figure 12), which is typical of a limited assembly of exactly parallel helices, the relative height of the packing peak, when expressed on the same basis as is used in table 3, is 0·92.

The position of the packing peak varies considerably, and there is no reason to doubt that this reflects differences in the mean distance of approach of neighbouring helices. Thus, the highest d value is given by poly-γ-benzyl-L-glutamate, with its exceptionally bulky side-chains, and the next highest by the equivalent methyl derivative. Six proteins exhibit identical spacings at $9{\cdot}8_4$ Å, while those for pepsin and β-lactoglobulin appear to be unusually high. Edestin is the only substance that gives a double peak, one spacing being exceptionally low and the other rather high. It is possible in this case that two distinct molecular, or submolecular, species are involved.

It is of interest that the haemoglobin and myoglobin of horse give practically identical packing peaks as regards both position and height; this may be held to indicate a close structural relationship between these two proteins, as has already been suggested by

Kendrew (1950) on more substantial grounds. Ox haemoglobin is again closely similar to that of horse, but human haemoglobin gives a significantly lower peak spacing at $9{\cdot}5_3$ Å, as compared with $9{\cdot}8_4$ Å for the other two haemoglobins and for myoglobin. Human fibrinogen, on the other hand, gives the $9{\cdot}8_4$ Å spacing.

The translation of these apparent Bragg spacings into inter-helix distances cannot be a precise matter as the position of the packing peak depends upon the nature of the assembly and can only truly be related to inter-helix distances, which may have several values, by the appropriate use of equation (18). It can, however, be asserted on theoretical grounds that, for assemblies of parallel thin cylinders of radius 2 Å separated by axis-to-axis distances of *ca.* 10 Å, the apparent d value will be somewhat smaller than the inter-helix distance if the cylinders are several in number and grouped together compactly.

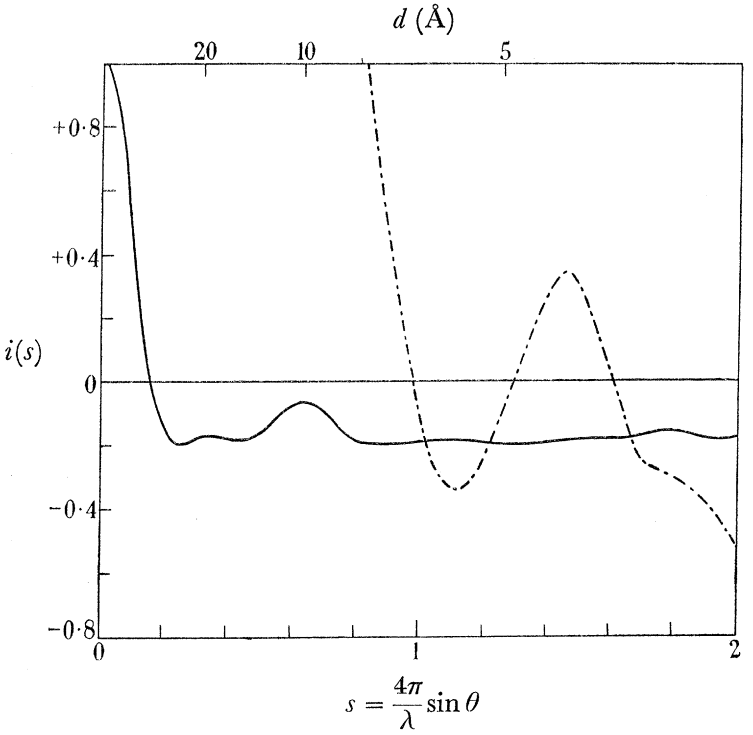

FIGURE 12. Inter-helix interference intensity curve for a centred hexagonal assembly of seven helices 11 Å apart (solid line). Broken line represents initial part of $i(s)$ curve for a single α_1-helix on the same scale. Note 'packing' peak of spacing 10 Å.

A likely example may be used to illustrate this point and to show graphically the nature of the packing effect. If seven cylindrical shells, each of radius 2 Å, be grouped together so that their axes form a centred regular hexagon of side 11 Å, the intensity curve for the assembly is easily calculated via equation (18) and takes the place of the rapidly rising intensity below $s \simeq 1{\cdot}0$ given by single α-helices in isolation. This is illustrated, in terms of $i(s)$, in figure 12.

It will be noted that the 'packing' peak has an apparent Bragg spacing of 10 Å, although the actual inter-helix distance in the assembly is 11 Å. The smaller subsidiary low-angle peak at *ca.* 18 Å that is also present, corresponding to second-nearest neighbour distances, is indicative of the limited lateral extent of the assembly for it cannot occur in an infinite

array. This peak is, in fact, observed experimentally with several proteins, as will be discussed in a subsequent paper describing low-angle interferences, principally of an inter-molecular character (by *molecule* is here meant the complete assembly of several helices). Figure 12 should be regarded in the following way. Below $s \simeq 1{\cdot}0$, the appropriate $i(s)$ curve is that shown with a continuous line and refers to inter-helix interferences; at higher s values, the appropriate curve is that representing intra-helix effects and is shown with a broken line. In practice, of course, the curves due to the two types of interferences merge smoothly into one another at $s \simeq 1{\cdot}0$.

An arrangement similar to the example just chosen is likely to occur in those proteins where a spatially economical packing of several helices is involved. For this type of system, no serious error is introduced if the apparent d value is multiplied by a factor of approximately 1·1 to give the mean axis-to-axis distance separating neighbouring helices. In this event, the frequently occurring spacing of $9{\cdot}8_4$ Å, for instance, would correspond to an average inter-helix distance of about 10·8 Å.

10. Conclusions

We have avoided other than parenthetical references to other investigations, as it was our intention to develop, as far as possible, a self-sufficient argument based solely on the evidence presented in this paper. Certain of the conclusions we have reached are, of course, greatly strengthened if account be taken of the existing corpus of knowledge regarding the structure of proteins and polypeptides. It has long been known, for instance, that single crystals of certain proteins possess marked concentrations of scattering matter separated by *ca.* 10 Å, and it has been considered highly probable in several cases that this corresponds to a parallel (or nearly parallel) grouping of rod-like subunits within the molecules (see, for example, Perutz 1949; Carlisle & Scouloudi 1951). The nature of the chain configuration within these rods has not, however, hitherto been determined except possibly for haemoglobin (Pauling & Corey 1951*c*; Perutz 1951). In respect of several proteins and synthetic polypeptides obtainable as fibres, strong evidence already exists for their possessing an α-helix configuration, and it is known that certain of them can also adopt a β-structure.

The main significance of the present work is that it generalizes these conclusions to apply to some twenty different corpuscular proteins (table 2, group α), if distinction is made between similar molecules from different animal species. In all these cases, it would appear, if our reasoning is valid, that the basic structure is the same in the sense that the *predominant* chain configuration is probably that of the Pauling & Corey α_1-helix and that the complete molecule essentially consists of a number of these helices packed together compactly. It is improbable that any other regular configuration occurs in linear segments of appreciable length as a *substantial* proportion of the chain structure.

It is, of course, possible that alternative configurations exist over limited segments of the chain or as separate chains in small amounts; in particular, the way in which a long α-helix bends back on itself is not defined. The closely related α_2-helix configuration is a special case; this could occur as an important proportion of the whole molecule although it is unlikely to be the major constituent. Our earlier suggestion (Arndt & Riley 1953) that

the insulin molecule is made up of equal numbers of α_1- and α_2-helices remains hypothetical and has not been capable either of substantiation or disproof in the present work. From this standpoint, it is unfortunate that the configurations in the separated *A* and *B* fractions obviously differ from those in the native protein molecule, but the observation in itself provides an interesting example of an $\alpha \rightarrow \beta$ transformation. In general, β-structures seem to occur less frequently than α.

While it is true that every conceivable configurational hypothesis has not been examined, enough possibilities have been considered to make it unlikely that one can exist that simulates in detail the scattering characteristics of the α-helix while differing from it sufficiently to be worthy of separate identification. It does not follow that some variations in the defining parameters of the α-helix would be unacceptable or that the helix need be perfectly regular over its whole length; considerable deformation is, however, in our view, excluded.

If the foregoing conclusions are accepted, it would be reasonable to extend them by analogy to all corpuscular proteins and to assume that a common basic molecular architecture is shared by all. The natural configuration for corpuscular proteins is α, and β is either exceedingly rare or is artificially induced.

We wish to express our indebtedness to Mr W. A. Coates and Miss Mary Tuck for their assistance during these researches. We are grateful to Professor L. Pauling, For.Mem.R.S., for the interest he has shown in this work and the encouragement he has given us.

Our sincere thanks are due to all those who have provided us with a variety of samples for examination and have thus helped to extend the scope of the investigations.

The work reported in this paper formed part of a program of research to which generous support was given over a period of several years by the British Empire Cancer Campaign; it was carried out in the Davy Faraday Laboratory of the Royal Institution, and our thanks are due to the Managers for facilities provided.

References

Arndt, U. W., Coates, W. A. & Crathorn, A. R. 1954 *Proc. Phys. Soc.* B, **67**, 357.
Arndt, U. W., Coates, W. A. & Riley, D. P. 1953 *Proc. Phys. Soc.* B, **66**, 1009.
Arndt, U. W. & Riley, D. P. 1952 *Proc. Phys. Soc.* A, **65**, 74.
Arndt, U. W. & Riley, D. P. 1953 *Nature, Lond.*, **172**, 245.
Astbury, W. T. 1931 *Ann. Rep. Chem. Soc., Lond.*, **28**, 322.
Astbury, W. T. 1940 *J. Int. Soc. Leath. Chem.* **24**, 69.
Astbury, W. T. 1953 *Proc. Roy. Soc.* B, **141**, 1.
Beevers, C. A. 1952 *Acta Cryst.* **5**, 670.
Bernal, J. D. & Crowfoot, D. 1934 *Nature, Lond.*, **133**, 794.
Bjørnhaug, A., Ellefsen, Ø. & Tønnesen, B. A. 1954 *J. Polym. Sci.* **12**, 621.
Bragg, W. L., Kendrew, J. C. & Perutz, M. F. 1950 *Proc. Roy. Soc.* A, **203**, 322.
Carlisle, C. H. & Scouloudi, H. 1951 *Proc. Roy. Soc.* A, **207**, 496.
Debye, P. 1915 *Ann. Phys., Lpz.*, **46**, 809.
Debye, P. 1930 *Phys. Z.* **31**, 419.
Donohue, J. 1953 *Proc. Nat. Acad. Sci., Wash.*, **39**, 47 0.
James, R. W. 1932 *Phys. Z.* **33**, 737.
Kendrew, J. C. 1950 *Proc. Roy. Soc.* A, **201**, 62.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Low, B. W. & Grenville-Wells, H. J. 1953 *Proc. Nat. Acad. Sci., Wash.*, **39**, 785.
Meyer, K. H. & Mark, H. 1928 *Ber. dtsch. chem. Ges.* **61**, 1932.
Osborne, T. B. 1902 *J. Amer. Chem. Soc.* **24**, 140.
Oster, G. & Riley, D. P. 1952 *Acta Cryst.* **5**, 272.
Patterson, A. L. 1944 *Phys. Rev.* **56**, 195.
Pauling, L. & Brockway, L. O. 1934 *J. Chem. Phys.* **2**, 867.
Pauling, L. & Corey, R. B. 1951*a* *Proc. Nat. Acad. Sci., Wash.*, **37**, 235.
Pauling, L. & Corey, R. B. 1951*b* *Proc. Nat. Acad. Sci., Wash.*, **37**, 729.
Pauling, L. & Corey, R. B. 1951*c* *Proc. Nat. Acad. Sci., Wash.*, **37**, 282.
Pauling, L. & Corey, R. B. 1953 *Proc. Nat. Acad. Sci., Wash.*, **39**, 253.
Pauling, L., Corey, R. B. & Branson, H. R. 1951 *Proc. Nat. Acad. Sci., Wash.*, **37**, 205.
Perutz, M. F. 1949 *Proc. Roy. Soc.* A, **195**, 474.
Perutz, M. F. 1951 *Nature, Lond.*, **167**, 1053.
Riley, D. P. & Arndt, U. W. 1952 *Nature, Lond.*, **169**, 138.
Riley, D. P. & Arndt, U. W. 1954 *J. Colloid. Sci.* (Supp.) **1**, 57.
Sanger, F. 1949 *Biochem. J.* **44**, 126.
Titchmarsh, E. C. 1937 *Introduction to the theory of Fourier integrals*. Oxford: The Clarendon Press.
Warren, B. E. & Gingrich, N. S. 1934 *Phys. Rev.* **46**, 368.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES